

DOCUMENT RESUME

ED 196 931

TM 810 043

AUTHOR Lennon, Roger T.
TITLE The Anatomy of a Scholastic Aptitude Test.
INSTITUTION National Council on Measurement in Education,
Washington, D.C.
PUB DATE 80
NOTE 9p.
AVAILABLE FROM National Council on Measurement in Education, USPS
823120, 1230 17th Street N.W., Washington, DC 20036
(\$2.50).
JOURNAL CIT NCME Measurement in Education: v11 n2 Win 1980
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Aptitude; *Aptitude Tests; Standardized
Tests; Test Construction; Test Interpretation; Test
Reliability; *Test Results; Test Validity

ABSTRACT

A brief, nontechnical state-of-the-art review of four common instruments and current practices in the field of scholastic aptitude testing is presented. The instruments include: the 1971 Cognitive Abilities Test (CAT), 1973 Henmon-Nelson Tests of Mental Ability (H-NTMA), 1979 Otis-Lennon School Ability Test (OLSAT) and the 1970 Short Form Test of Academic Aptitude (SFTAA). The fundamental premise underlying the use of these tests is that pupils vary with respect to possessing the abilities needed to successfully master the academic learning tasks required by schools. Knowledge about these abilities is useful for planning instruction. Reexaminations of original premises that had led both to the development and the use of the tests has resulted in modifications, both with respect to what the tests purport to do and with respect to their development and interpretation. Of the four tests, only H-NTMA continues to identify itself as a test of "mental ability." The others are presented as tests of "school ability," "cognitive abilities," and "academic aptitude." The present designations reflect more modest representations than previous tests. Topics discussed include interpreting scores, test development and standardization, validity, reliability, and using test results. (Author/RL)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

the editor

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRO-
DUCED EXACTLY AS RECEIVED FROM
THE PERSON OR ORGANIZATION ORIGIN-
ATING IT. POINTS OF VIEW OR OPINIONS
STATED DO NOT NECESSARILY REPRESENT
OFFICIAL NATIONAL INSTITUTE OF
EDUCATION POSITION OR POLICY.

Vol. 11, No. 2

Winter, 1980

The Anatomy of a Scholastic Aptitude Test

Roger T. Lennon

ED196931

ABOUT THIS ISSUE



figure in educational measurement in the United States and abroad for more than three decades. His ability to analyze issues fairly and with a high degree of competence has earned him the covetous respect of his peers.

We welcome back, once again, Dr. Roger T. Lennon to the pages of *ME*. When this editor identified the need for a close look at the anatomy of scholastic aptitude tests he had to look for an author no further than the current president of the National Council on Measurement in Education. Roger has been a prominent

Group intelligence tests, scholastic aptitude measures, school ability measures — by whatever name they are called — have been the targets of litigation, criticism and, at times, villification of the vilest sort. What are these tests? What do they really measure? Are the criticisms justified? In order to add to the on-going dialogue — albeit in a constructive manner — *ME* asked Dr. Lennon to prepare an analysis of the type of test more commonly known as group-administered "intelligence tests." *ME* supplied the title and Dr. Lennon supplied the content; from an editor's point-of-view, a reasonable division of labor.

Dr. Roger T. Lennon currently serves as Associate to the Chairman in Measurement, Harcourt Brace Jovanovich, Incorporated, and is the author of the *Otis-Lennon School Ability Test*.

HCR

From the earliest days of use of standardized tests in schools, the *group intelligence* or *mental ability* test has been among the most widely used, if not the most widely used, of such examinations. Indeed, the lay person's concept of a standardized test is most likely to be of a group intelligence test — an "IQ test"; he is even likely to refer to any standardized test as an "IQ test." Most readers of *Measurement in Education* may be presumed to be familiar with common instruments and practices in this field of testing, both through exposure to them in the course of their formal training, and through direct involvement in the field. At the same time, it may be useful even for such informed readers to have a brief, non-technical state-of-the-art review, or update. This type of testing is increasingly controversial, and it is interesting to see how the current versions of typical instruments reflect

test authors' and publishers' responses to frequently voiced concerns.

Group intelligence or mental ability tests, never free of some criticism, have in the past thirty years been the

NOTICE: Due to a regulation of the United States Post Office the next issue will begin Volume 12, No. 1. Readers will continue to receive the number of issues they should receive as members of NCME. This will be done to match the volume and issue numbers with the calendar year.

TM 810043

objects of intense scrutiny, often hostile, by educators and lay persons. They have been accused of being unfair to members of minority subgroups in the population, of fostering lowered expectations and thus reduced educational opportunities for persons who score poorly on them, of fostering inaccurate or irrelevant notions concerning the nature of ability, of providing no useful information about learners, and of other defects. The merits of these objections have been discussed by the writer in an earlier issue of *Measurement in Education* (Spring, 1978), and it is not the purpose here to extend this debate. Rather, this article seeks to describe the character of several current exemplars of this type of instrument, their presumed role in the school, how the current editions differ from earlier ones, the evidences that these tests do serve the intended purposes, and the manner in which they are constructed to serve these ends.

We have chosen to examine four instruments in this category, all of which enjoy widespread use in American schools, all of which trace their origins back through several predecessor editions, and all of which have been perceived by reviewers as sound instruments psychometrically. The tests are the *Cognitive Abilities Test* (1971),* *Henmon-Nelson Tests of Mental Ability* (1973), *Otis-Lennon School Ability Test* (1979), and the *Short Form Test of Academic Aptitude* (1970). Collectively, these four instruments would account for a very large fraction of tests of this kind used in schools and are surely representative of this class of tests. Although there are differences in content, structure and organization among them, there is much in common among them, conceptually and empirically, and they are regarded as competitive — serving the same purposes — by users.

The fundamental premise underlying the use of these tests in the school is that pupils vary importantly with respect to their possession of the various abilities which are important for success in mastering the academic learning tasks of the school. Knowledge about these abilities is useful — (1) for the teacher, in planning instruction appropriate for individual pupils, (2) for school administrators, for purposes of curriculum planning, grouping, selection of materials, and the like, and (3) for guidance counselors, in order better to counsel students with respect to educational and career planning. There are other contexts in which this knowledge is judged to be useful, and even essential, as in research studies where it is necessary to control for pupil learning ability, or where it is desired to assess the effects of various educational programs, but this article will concentrate more on the former set of purposes.

The historical roots of this type of test are well-known. They are to be found in the early twentieth-century efforts of Alfred Binet to devise a means for identifying pupils who would not succeed with the ordinary work in the Paris schools. His efforts eventuated in the *Binet-Simon Scales*. This work was adapted for use in America by several investigators, the most prominent of whom was Lewis Terman, through his Stanford Revision of the Binet Scales. One of Terman's students, Arthur Otis, undertook to develop a version of the Binet test suitable for administration to groups of examinees at a time; Otis' work was embodied in the *Army Alpha Tests* of World War I and formed the basis for his *Otis Group Intelligence*

Scale, the first of this type of tests to achieve widespread use in American schools in the early 1920's. There quickly followed in the 1920's dozens of published tests of this type. Textbooks were devoted to the development, use, and interpretation of such measures, and the professional literature abounded with discussions of them.

The controversy over the nature of intelligence — the g-theory versus factor theories — was reflected in the variety of tests developed and published, but it continued to be assumed that some measure of intelligence, however it might be defined, had a useful place in education. The nurture issue — the susceptibility of performance on these tests to environmental influences — was pursued with vigor in the '30's but, curiously, did not provoke questions about bias in the tests, at least to anything like the extent that such questions were raised beginning in the late 1940's.

NCAE

measurement in education

Vol. 11, No. 2 Winter, 1980

Editor: Herbert C. Rudman, Editor

Editorial Advisory Board:

Joan Benbenbaker, Cincinnati Public Schools
 William C. Coffman, University of Iowa
 Barry Guttman, Saginaw Public Schools
 James C. Sanders, Western Michigan University

Measurement in Education (USPS 023120) is a series of special reports published four times a year in October, January, March and May by the National Council on Measurement in Education. These reports are concerned with the practical implications of measurement and related research and their application to education: problems of individuals, institutions and systems. The emphasis is upon uses of measurement rather than technical or theoretical issues. Subscription rate: \$6.00 per year; single copies 75¢ each in quantities of 25 or more, or \$2.50 for a single issue.

Second-class postage paid at Washington, D.C.
and additional mailing points

Address Editorial correspondence to:

Dr. Herbert C. Rudman
 Room 405, College of Education
 Michigan State University
 East Lansing, MI 48824

Address business correspondence to:

Dr. William J. Russell
 123 17th Street, N.W.
 Washington, DC 20036

*A revised edition of *Cognitive Abilities Test* was published in 1978. The conclusions in this article, though based on the 1971 edition, are still in general appropriate.

Use of these tests continued unabated following World War II, despite increasing disposition to question their utility. By the late 1960's, however, questions of cultural bias in tests had become so insistent, as had questions concerning the consequences of the use of this type of test in schools, as to prompt reexaminations of the original premises that had led both to the development and the use of the instruments. Results of these reexaminations are apparent in current editions of the tests, both with respect to what the tests purport to do and with respect to their development and interpretation.

Intelligence or Scholastic Aptitude?

Of the four tests, only *Henmon-Nelson* continues to identify itself as a test of "mental ability." The others in their current editions are presented as tests of "school ability," "cognitive abilities," and "academic aptitude." These designations are significant: the latter three tests in earlier editions had more ambitious designations — *mental ability*, *intelligence*, and *mental maturity*, respectively. The present designations reflect more modest representations of what the tests measure, or at least the purposes for which they are most recommended. Given the relatively minor departures in content and structure from earlier editions, it will occur to many to wonder whether the tests should still be considered "intelligence" tests, as were their predecessor editions and, if not, what different conceptions prevail.

For an answer to this question, it is useful to go to the declarations in the respective manuals for these tests. Here we find the following:

"... the series was designed to provide an accurate measure of the abilities needed to acquire the desired cognitive outcomes of formal education. This complex of abilities, sometimes designated 'scholastic aptitude,' is what is intended by the term 'school ability' as used here. ...

"For a large majority of examinees, results are indicative of learning abilities in situations other than the school context. However, since some examinees have sufficiently different experiential backgrounds and opportunities to learn the types of tasks, it is unwise to suggest broader interpretation." (OLSAT, 1979)

"... a set of measures of the individual's ability to use and manipulate abstract and symbolic relationships. Three main types of symbols play substantial roles in the thinking of students and adults: symbols representing words, symbols representing quantities, and symbols representing spatial, geometric or figural patterns.

"... the authors have attempted to emphasize relational thinking — the perceiving of relationships among abstract elements in a variety of media and settings. Throughout each of the subtests the basic elements have been kept relatively simple, clear and familiar. The intended emphasis is on discovery of relationships and flexibility of thinking. Exotic, specialized, or trivial content has been avoided.

"... the test appraises developed abilities. These abilities represent the interaction of a life history of experiences impinging upon a specific biological organism." (CAT, 1971)

"... was developed to measure the results of these influences (educational experience, biological attributes, cultural experiences, and maturational rate) upon scholastic potential; ... Nor does the test measure aptitude for many things outside of school. ... Probably the test could suggest a general level of intelligence for dealing with problems outside of school, but the symbols and concepts it uses are so academic in nature that to generalize beyond school is generally unwarranted." (SFTAA, 1970)

"... to measure those aspects of mental ability which are important for success in school work. Scores may be interpreted as measures of general mental ability or 'academic' aptitude. High performance requires the efficient utilization of verbal and numerical symbols and the ability to acquire and retain information in common symbol form for use at later times." (H-NTMA, 1973)

From the foregoing, as well as from a more intensive study of the manuals for the four tests, it may be judged that there is a certain equivocation or ambivalence on the part of their authors as to whether the tests continue to be "intelligence" or "mental ability" tests, or should be regarded only as measures of school learning ability. The representations in the current manuals are notably more cautious and limited with respect to the abilities measured. *The tests, in a word, are proffered as measures of those abilities that are important for mastering the cognitive tasks set by the school; the evidences of validity and the recommended modes of interpretation and use are consistent with these avowed purposes.* If some wider significance is to be attributed to the scores, the authors seem to be saying, it must be established independently and goes beyond the more common purposes of the tests.

There is general rejection in the manuals for all the tests that what is measured is innate, fixed, constitutional, or not influenced by environmental forces. Thus, for example,

"... scores should never be regarded as measures of a presumed innate or congenital capacity. They are the results of complex and interacting influences that probably include genetic elements, but which most certainly include environmental and experiential elements." (OLSAT, 1979)

"Although knowledge is required to answer the items, they were selected to reflect ability to profit from exposure to learning in a variety of situations." (H-NTMA, 1973)

The Content of Scholastic Aptitude Tests

Although these tests now hold themselves out as assaying "scholastic aptitude" rather than *intelligence* or *mental ability*, as in earlier editions, they continue to rely on much the same types of content as they did earlier, in the belief that these types of test content are effective in

predicting achievement in scholastic areas. Thus, all tests include elements that may be designated *verbal* — vocabulary, opposites, sentence completion; all include *reasoning* items — analogies (verbal and figure), classification, number and/or letter series, and inference; *quantitative*, including arithmetic reasoning and quantitative concepts; and, in some cases, *information* and *memory* items. The item types are intended to be ones on which success will not be predominantly dependent upon in-school instruction but will rather reflect learnings that have come about through the general life experience of the examinees. They are intended to reflect learnings that have arisen through experiences that are common to the vast majority of examinees, rather than learnings consequent upon specialized or rarely occurring experiences. Given the more sharply-defined focus of the tests, viz. the prediction of success in learning academic material, the subtests are chosen with particular reference to their presumed effectiveness in manifesting ability to do abstract thinking, or relational thinking, and ability to manipulate verbal and numerical symbol systems.

The test authors have been attentive to criticisms that their instruments are “just other achievement tests,” dependent on results of in-school instruction, and deserving of no more acceptance as measures of “capacity,” “potential,” or “aptitude,” than tests clearly recognized as “achievement” tests — reading, language, etc. They recognize that the behaviors sampled by the tests are, as they must be, *learned behaviors, developed abilities*; and in this sense the tests are achievement measures. The authors, however, are at some pains to make clear that what their tests measure is not achievement explicitly sought through school experience. Witness declarations such as the following:

“... (The test) does not measure how much a student has learned in school. The abilities this test measures are those which fledge out indigenously by simple participation in the practical workaday world of most persons. What students have learned in school might be influential, but only slightly so.” (SFTAA, 1970)

“Scores may be interpreted as measures of general mental ability, or academic aptitude, but not as indices of school achievement specifically taught in a common curriculum.” (H-NTMA, 1973)

Interpreting Scores

There is considerable uniformity among the four tests in the arrangements they provide for interpretation of scores. In the case of all four tests, scores may be converted to a system of normalized standard scores in which the average score for the reference group — an unselected group, either single-age or all ages combined — is 100 and the standard deviation 16. These values will be recognized as the defining constants for the familiar IQ — so much so that the derived scores, in the case of two of the tests in question, are called “deviation IQ’s,” even though they are not quotients in any sense of the word, and even though the respective authors do not represent the tests as “intelligence” tests, with only the word “deviation” signalling the manner of their derivation. The interpreted score is designated *Standard Age Score* in the

case of one of the other tests, and *School Ability Index* for the other.

For purposes of convenience, all of the interpreted scores will be referred to in this discussion as deviation IQ’s (DIQ’s). Their derivation is, for all practical purposes, the same in all four measures, involving a transformation of the raw score distributions for single-age populations to normalized distributions having the designated constants. The normalized distributions of DIQ’s are similar from age to age, thus incurring uniformity of meaning across the age range. The test authors seem desirous of discouraging the use of IQ as the interpreted score, but retain it, even though subordinated to other modes of interpretation, in the belief that many users feel more comfortable continuing to use a terminology for which they have built up a body of understanding. A strong case may be made for abandoning the term “IQ” and perhaps the metric itself, even when derived according to the deviation method. Terms such as *Standard Age Score* or *School Ability Index* are free of the objectionable connotations which many critics feel attend use of the term IQ.

As alternative interpretations to IQ’s, all these tests provide systems of percentile ranks and stanines for age and grade groups. Compared with earlier editions of most of the tests, the current versions play down, to the point of elimination in some cases, *mental ages*, or other power measures.

The DIQ’s, or the percentile ranks or stanines, provide a measure of the examinee’s relative standing in ability to do schoolwork among a representative group of examinees of similar age or grade. In the case of all tests, ways are presented of relating these interpreted scores to measures of achievement and of utilizing them in connection with adaptation of the instructional program to the apparent ability level of the examinee.

Development and Standardization

The development process for the four tests followed a similar pattern. After the identification of the abilities to be measured and the types of items best calculated to measure them, a pool of items was created, drawn in some instances from items in previous editions but in most cases newly created for the current editions, and numbering three to four times as many items as would be required for the published forms. In most cases this pool of items was reviewed by representatives of minority groups, in an attempt to identify items which on inspection would appear to have any biases against these groups.

All items were then subjected to experimental tryout in administrations to large and presumably typical samples of American school pupils. On the basis of these tryouts, difficulty and discrimination indices were computed for all items; in some cases, statistical analyses were made of the performance of the majority and minority subgroups on the items as an additional technique for identifying possibly biased items. Where such evidences of possible bias arose, the items in question were again subjected to review by minority representatives and any items for which a substantial argument could be made for supposing bias existed were eliminated.

Final forms of the various levels of the tests were

constructed so as to yield measures of specified range of difficulty and discriminating power. Appropriate methods were applied to establish comparability among successive levels and to verify the equivalence of alternate forms.

For all four tests the test construction processes followed, as described in their respective manuals, represent conscientious craftsmanship, and are particularly noteworthy for their care in eliminating potential sources of bias.

The procedures for developing norms for the several tests also have a great deal in common. In all cases the authors and publishers sought to develop norms that would accurately represent, at the time of the respective standardization, the performance of the American school population on the various instruments. The attainment of this goal was sought through the identification of samples of school systems, including public and non-public, that could be considered representative of the total school population. In the case of all tests, several stratifying variables were used to help achieve desired representativeness; region, size and type of system, and socio-economic status of the area served by the participating school district. Research has indicated that these variables are related to performance on the type of test in question, and that in combination they account for a large part of the variance in test scores identified with available demographic indices. In certain cases where the obtained normative sample did not conform precisely to the model defined by census or other data, appropriate weightings were applied to data in the various strata to make the sample match the model more precisely. The bodies of normative data developed through these standardization programs are purposeful, well-defined, and deserving of serious consideration as descriptive of the performance, at the times in question, of the American school population. Refinements from one standardization population to another, as with respect to ethnicity or with respect to treatment of special education groups, are judged to have very slight over-all impact on the final norms.

In the case of all four tests, the results for the examinees included in the normative samples were summarized for age and grade groups using, where necessary, the score conversions among levels described above. There resulted distributions of scores for examinees of given age or grade status; these distributions formed the basis for conversion of scores to DIQ's, percentile ranks or stanines within age or grade groups.

While the care exercised in the standardization of the several tests does not necessarily create direct comparability among their results (their respective standardizations took place over a ten-year period), the various norming programs embody such consistent technology and carried it out so carefully in each case, as to make it more probable than would have been true thirty years ago.

Validity

The validity of the tests is addressed in their respective manuals and other supplementary technical publications under the familiar categories of *content validity*, *construct validity* (termed "congruent validity" in the case of one of the tests), and *criterion-related validity*, expressed in both *concurrent* and *predictive* senses.

Content validity

The case for content validity typically consists of a brief designation of the domain presumably being sampled by the test content — e.g., "facility in dealing with abstract concepts presented in verbal, figural, or symbolic form... ability to discern relationships of varying complexity"; "general abstract reasoning ability"; "relational thinking"; etc.; brief descriptions of the types of behaviors making up the domain, and occasionally an effort to specify the critical dimensions of the domain. Discriminating power of items finally included in the tests, and in some instances inter-correlations among item types used, are sometimes cited in support of content validity claims. The discussions of content validity tend to be relatively brief and to rest on rational considerations rather than empirical data. There is an appeal to tradition in most of the discussions, citing the fact that most of the types of items employed have had a long history of use in tests of this kind. One of the manuals refers to the content as made up of types of items which have "stood the test of time."

Construct validity

Discussions of construct validity tend to be brief. In some cases, data on intercorrelations among the parts of a test are offered as evidences of construct validity, as in one instance is a factor analysis of scores on parts of the test which is reported to reveal a "strong, common factor," presumed to correspond with the ability being measured. Other evidences of construct validity include correlations of the scores on the various tests with scores on other tests purporting to measure the same or a similar ability. For example, correlations are reported between some of the tests under discussion and *Stanford-Binet* results, *Differential Aptitude Test* results, *Raven Progressive Matrices* scores and the like. Such correlations typically are moderate (.50 to .60), but are presumably supportive of the proposition that the tests share common sources of variance with other tests intended to serve similar purposes.

Criterion-related validity

By far the most impressive evidences of validity are those pertaining to criterion-related validity. Given the avowed purposes, and the most common uses of these tests, this is the type of validity evidence that is most appropriate. For these four tests collectively, there is a wealth of information on correlations of the scores which they yield (including part scores and total scores) with measures of achievement, in the form of either teacher grades or achievement test scores. These data include correlations between scholastic aptitude and achievement measures obtained either concurrently or considerably later — six months, a year, two, or even more years; truly predictive data. These correlations are uniformly positive and of moderate to appreciable magnitude. They tend to increase the higher one goes up through the grades. In grades four and above, the correlations between scores on any of these tests and achievement, as measured by any of the most widely-used achievement batteries, tend to cluster in the .70 to .85 range. Correlation with reading test scores tend to be highest,

followed closely by mathematics and language, and lowest with spelling. Correlations of scholastic aptitude scores with teacher grades are somewhat lower than with achievement test scores, but still appreciable.

Correlations with achievement are so high as to seem to lend credence to the sometimes heard complaint that the scholastic aptitude tests are really measuring the same thing as the achievement tests and that there is, therefore, redundancy in administering both types of tests. To this it may be replied that even with a correlation of .80 — and even making allowance for the unreliability in both measures — there is still a degree of independent variance which will result in the identification of some fraction, admittedly small, of pupils with significant discrepancies in the two types of score. One need not be surprised that the majority of youngsters identified as high in scholastic aptitude do well in the various branches of the school curriculum, and vice versa. To discover otherwise would call into serious question the validity of the scholastic aptitude measure. However, by the intent of the test makers, the tasks in the scholastic aptitude tests are only minimally influenced by specific in-school instruction; use of the two types of test permits, therefore, a comparison of a child's success in school learning with his success in learnings more dependent on general life experiences; surely a useful comparison in certain efforts at diagnosis of learning problems, evaluation of the effectiveness of instructional efforts, and planning appropriate interventions.

Reliability

In the case of all four tests, ample data are provided concerning the reliability of the results. These reliability data include Kuder-Richardson 20, test-retest, alternate-form, and odd-even values, not every type of information in the case of every test, but on the whole adequate to permit confident declarations about the stability of the test results. It is a fair generalization that in all cases, total scores or major sub-scores such as verbal and non-verbal, language and non-language, meet textbook standards of reliability for the purposes for which the tests are recommended. Typical K-R 20 or test-retest values for total scores are above .90; very few reliability estimates for any grade, even for part scores, fall below .85. Reliabilities of these scores compare favorably with those of either individual intelligence tests or of even the best of achievement measures. In the case of all tests, data are also provided on standard error of measurement, so that the user can estimate the confidence to be reposed in a given score. These data indicate that the error involved in a score is within a tolerable range. Although the data are presented sometimes in terms of raw scores, sometimes in terms of other reference scores, they translate into a standard error of measurement for DIQ's of the order of 5-6 points.

Some of the tests also provided information about stability of test results over longer time periods; up to as much as nine years. The correlations over extended periods, and the similarity of means and standard deviations resulting from administrations separated by considerable time, make clear that the tests are measuring a characteristic of substantial persistence and consistency;

CAT (1971) for example, reports correlations between administrations three years apart averaging .60 between grades 3 and 6 and somewhat higher between grade 6 and 9. Similarly high agreement has been reported for Otis-Lennon (1979) results from administrations at three and even six-year intervals.

Using the Test Results

Recommended uses of the results of these tests may readily be inferred from what has been said about their nature, content and development. Knowledge of how learners stand in school learning ability is judged to be important, if not critical, in establishing an appropriate instructional program — appropriate as to level of content, manner of presentation, and pace. Interpretation of the learner's performance in the various achievement areas, and particularly diagnosis of difficulties when the performance is judged to be unsatisfactory, is facilitated through information about the school learning ability brought to the task.

The manual for one of the tests (CAT, 1971) enumerates possible uses of test results as follows:

1. Selecting curriculum materials
2. Organizing small work groups
3. Setting the pace of instruction
4. Selecting methods of instruction
5. Differentiating goals and expected levels of achievement
6. Identifying disabilities
7. Counseling students
8. Reporting to parents
9. Surveying the school

These uses are echoed, implicitly or explicitly, in the manuals of the other tests. The listed uses constitute a reasonable catalog of situations in which information about learning ability, as the several tests measure it, will lead to more appropriate instructional practice.

It may be argued, and with some justice, that the manuals provide general but not specific guidance as to how the results may be used in the several contexts suggested. Nowhere, for example, is information set forth on exactly what curricular adjustments should be made for a child of Deviation IQ 85 in the 5th grade with respect to his arithmetic program; nor are there explicit rules relating scholastic aptitude to particular learning difficulties, such as would permit cookbook-like diagnoses of the nature of learning disabilities; nor are the connections between scholastic aptitude scores and pace of instruction anywhere made very definite. The reason for such lack of specificity is not hard to find; our knowledge base is not adequate to sustain highly specific declarations about the significance of given scholastic aptitude scores for the fine-tuning of instructional interventions. It does not follow, however, that because these matters cannot be spelled out with the precision and exactness that might be desired, the information is worthless. It must be remembered that with or without scholastic aptitude test results, teachers are constantly, and inescapably, making their own appraisals of learner capability and adjusting their instruction to such knowledge. The quantification of the estimate of scholastic aptitude makes available estimates

of this attribute of known dependability and known relation to achievement.

Certain of the tests have developed explicit schemes for relating achievement in the various subjects at the various grade levels to measured scholastic aptitude, which permit comparison of the two to be made, and determination of the existence of significant discrepancies between them. These schemes make possible the convenient identification of pupils whose performance in a given subject at a given grade level differs sufficiently from the performance of other pupils of similar scholastic aptitude to warrant special attention by the teacher.

The manuals also reflect a sensitivity to the questions that have, properly, been raised about the concepts of over- and under-achievement, and are careful to avoid the simplistic interpretations of differences between achievement and aptitude measures that have provoked criticisms. The manuals recognize the multiple influences on both achievement and aptitude scores that must be taken into account in seeking explanations of discrepancies between them, and offer generally cautious and conservative guides to such explanations.

The authors have been conscious of the criticism often

leveled at the use of scholastic aptitude measures to the effect that, by affecting teacher expectations, they may give rise to a diminution of educational opportunity. The argument goes that if teachers are led by low performance on scholastic aptitude tests to expect poor achievement from a pupil, they will consciously or unconsciously modify their instructional behavior, as by offering watered-down content presented at a slow pace, and thus contribute to fulfilling a prophecy of limited achievement. Attention is drawn to the wide range of achievement associated with any level of scholastic aptitude, even low levels, and the teacher is reminded that the task is to get from every pupil the maximum of which he or she is capable, and not to settle for the average level of achievement of pupils of the particular scholastic aptitude level. Recent research on the so-called expectancy phenomenon, incidentally, fails to support the apprehensions that have been voiced about the constriction of educational opportunities supposedly deriving from teachers' knowledge of limited scholastic aptitude.

The bodies of information collected about the relation of scholastic aptitude measures to achievement measures are essential in most studies of the effectiveness of

REPORTS AVAILABLE

Back issues of *Measurement in Education* are available for \$2.50 each per single issue, and for 75¢ each in quantities of 25 or more for a single issue.

- | | | | |
|---------------|--|----------------|---|
| Vol. 1, No. 1 | <i>Helping Teachers Use Tests</i> by Robert L. Thorndike | No. 4 | <i>Dear Mama: Why don't they love me anymore?</i> by Thomas J. Fitzgibbon |
| No. 2 | <i>Interpreting Achievement Profiles — Uses and Warnings</i> by Eric F. Gardner | Vol. 7, No. 1 | <i>The Discrepancy Evaluation Model (Part I)</i> by Andres Steinmetz |
| No. 3 | <i>Mastery Learning and Mastery Testing</i> by Samuel T. Mayo | No. 2 | <i>The Discrepancy Evaluation Model (Part II)</i> by Andres Steinmetz |
| No. 4 | <i>On Reporting Test Results to Community Groups</i> by Alden W. Badal & Edwin P. Larsen | No. 3 | <i>What Graduate and Professional School Students Think About Admission Tests</i> by Leonard L. Baird |
| Vol. 2, No. 1 | <i>National Assessment Says</i> by Frank B. Womer | No. 4 | <i>The Paradox of Education Testing</i> by Robert L. Ebel |
| No. 2 | <i>The PLAN System for Individualizing Education</i> by John C. Flanagan | Vol. 8, No. 1 | <i>FACE UP: A Framework for Assessing Career Educators Utilization Programs</i> by Bruce W. Tuckman |
| No. 3 | <i>Measurement Aspects of Performance Contracting</i> by Richard E. Schutz | No. 2 | <i>Grade Equivalent Scores by Gary Echternacht and If Not Grade Equivalent Scores — Then What?</i> by Jeanne M. Plas |
| No. 4 | <i>The History of Grading Practices</i> by Louise Witmer Cureton | No. 3 | <i>Criticisms of Testing: How Mean is the Median?</i> by Henry S. Dyer |
| Vol. 3, No. 1 | <i>Using Your Achievement Test Score Reports</i> by Edwin Gary Joselyn & Jack C. Merwin | No. 4 | <i>The Superintendent and Testing: Implications for the Curriculum</i> by Herbert C. Rudman |
| No. 2 | <i>An Item Analysis Service for Teachers</i> by Willard G. Warrington | Vol. 9, No. 1 | <i>A Consumer's Guide to Criterion-Referenced Test Item Statistics</i> by Ronald C. Berk |
| No. 3 | <i>On the Reliability of Ratings of Essay Examinations</i> by William E. Coffman | No. 2 | <i>Perspective on Intelligence Testing</i> , by Roger T. Lennon |
| No. 4 | <i>Criterion-Referenced Testing in the Classroom</i> by Peter W. Airasian and George F. Madaus | No. 3 | <i>Bias</i> , by Eric F. Gardner |
| Vol. 4, No. 1 | <i>Goals and Objectives in Planning and Evaluation: A Second Generation</i> by Victor W. Doherty and Walter E. Hathaway | No. 4 | <i>The Use of Tests in Admissions to Higher Education</i> by Mary Fruen |
| No. 2 | <i>Career Maturity</i> by John O. Crites | Vol. 10, No. 1 | <i>Some Comments on Criterion-Referenced and Norm-Referenced Achievement Tests</i> by William A. Mehrens and Robert L. Ebel |
| No. 3 | <i>Assessing Educational Achievement in the Affective Domain</i> by Ralph W. Tyler | No. 2 | <i>Does the Rasch Model Really Work? A Discussion for Practitioners</i> by R. Robert Rentz and Charlotte C. Rentz |
| No. 4 | <i>The National Test-Equating Study in Reading (The Anchor Test Study)</i> by Richard M. Jaeger | No. 3 | <i>A Perspective On The Uses and Misuses of Standardized Achievement Tests</i> by Peter W. Airasian |
| Vol. 5, No. 1 | <i>The Tangled Web</i> by Fred F. Harclerod | No. 4 | <i>The Practical Side of Criterion-Referenced Test Development</i> W. James Popham and Elaine Lindheim |
| No. 2 | <i>A Moratorium? What Kind?</i> by William E. Coffman | Vol. 11, No. 1 | <i>The Courts and Your Minimum Competency Testing Program — A Guide to Survival</i> by Thomas H. Fisher |
| No. 3 | <i>Evaluators, Educators, and the Publics: A Debate?</i> by William A. Mehrens | | |
| No. 4 | <i>Shall We Get Rid of Grades?</i> by Robert L. Ebel | | |
| Vol. 6, No. 1 | <i>On Evaluating a Project: Some Practical Suggestions</i> by John W. Wick | | |
| No. 2 | <i>Measuring Reading Achievement: A Case for Criterion-Referenced Testing and Accountability</i> by S. Jay Samuels and Glenace E. Edwall | | |
| No. 3 | <i>Security in a Citywide Testing Program</i> by Anthony J. Polemeni | | |

educational programs. In most such evaluative endeavors it is necessary to consider the quality of school learning ability which subject populations bring to whatever type of educational intervention is provided. Information about this school learning ability is essential if any assertions are to be made about the relative effectiveness of various types of educational programs for students of various levels of ability. Because the school learning ability scores do not reflect many of the out-of-school influences that have operated to produce variance in scholastic aptitude, they provide a convenient method for exercising at least partial control over these influences in studies of educational effectiveness.

In the case of certain of the tests under discussion, publishers provide services that relate achievement scores to scholastic aptitude scores for particular populations. The joint presentation of scholastic aptitude and achievement measures, sometimes in the form of bivariate plots between the two, provide clear portrayals of the degree of association between these measures, permit easy identification of conspicuous discrepancy cases, and, since they are based on local data, demonstrate the relationships prevailing in the local system rather than the more generalized relationship based on an entire standardization population.

The Future of Scholastic Aptitude Testing

Given the long history of the use of scholastic aptitude tests in American education, and their centrality in most established testing programs, it might seem beyond question to anticipate their continued widespread use for the foreseeable future. Indeed this would be true were it not for the fact that these tests have been the type most subject to criticism and attack during the past decade or more. They have been a particular target, for example, of

the anti-test advocates in the National Education Association, which has called for a moratorium on their use. The proscription against the use of group intelligence tests in New York City and in California has prompted some in other jurisdictions to reconsider the role of these tests in their programs. As far as available data indicate, these scholastic aptitude tests continue to enjoy about as much utilization as they ever did. Members of NCME, when surveyed a few years ago as to their opinion on trends in the use of these tests, were divided: 42% look for no change from the present, 42% predict less use, and 9% greater use.

The current editions of the four tests have clearly recognized some of the criticisms and taken measures to deal with them. The respective manuals present a more sophisticated view of what the tests are measuring, what types of inferences are supportable on the basis of the test scores, and what limitations need to be placed on interpretations of results. Teachers and others who study the manuals carefully should have little likelihood of making the kinds of mistakes in the use of these tests that gave rise to much of the criticism. If the use of this type of test diminishes, it will not be in response to criticisms or because of dissatisfaction. Rather, it is likely that they will gradually be replaced by other instruments capable of yielding more accurate or more efficient measures of abilities that are relevant to successful mastery of the academic tasks of the school.

References

- Lamke, Tom A., Martin J. Nelson, and Joseph L. French. *The Henmon-Nelson Tests of Mental Ability*. (H-NTMA) (Boston: Houghton-Mifflin Company, 1973)*.
- Otis, Arthur S. and Roger T. Lennon. *Otis-Lennon School Ability Test (OLSAT)*. (New York: Psychological Corporation, 1979).
- Sullivan, Elizabeth T., Willis W. Clark, and Ernest W. Tiegs. *Short Form Test of Academic Aptitude*. (SFTAA) (Monterey Park: CTB/McGraw-Hill, 1970).
- Thorndike, Robert L., Elizabeth Hagen, and Irving Lorge. *Cognitive Abilities Test*. (CAT) (Boston: Houghton-Mifflin Company, 1971)* (See footnote, page 2)

*Publisher is now Riverside Press.



NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION

USPS 823120

1230 17th Street, N.W.
Washington, DC 20036

Second class postage paid
at East Lansing, Mich.
48823 and
additional mailing points.

500003236 M
BARBARA M WILDEMUTH
TEST COLLECTION
EDUCATIONAL TESTING SERV
PRINCETON, NJ 08541